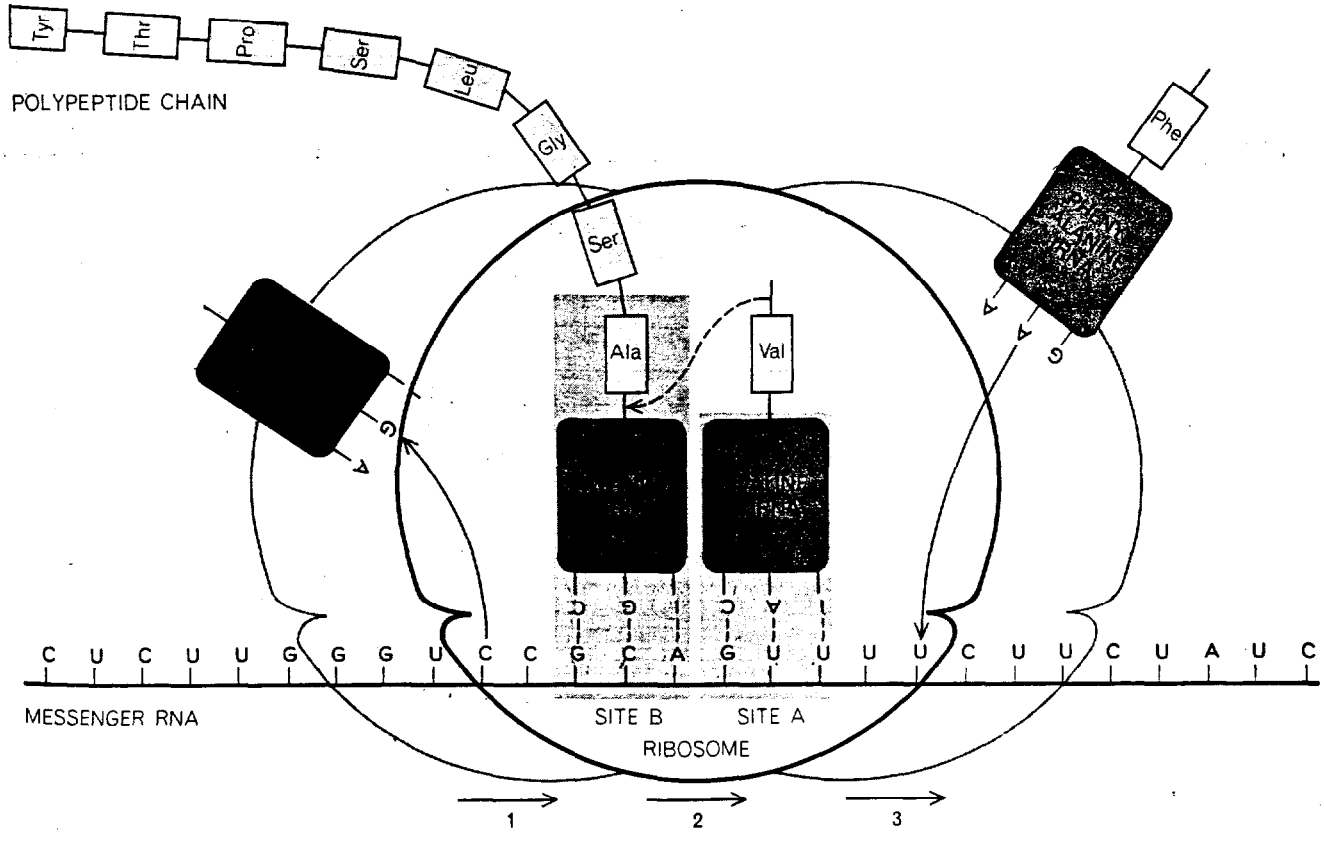
THE GENETIC CODE: III

The central theme of molecular biology is confirmed by detailed knowledge of how the four-letter language embodied in molecules of nucleic acid controls the 20-letter language of the proteins

by F. H. C. Crick

The hypothesis that the genes of the living cell contain all the information needed for the cell to reproduce itself is now more than 50 years old. Implicit in the hypothesis is the idea that the genes bear in coded form the detailed specifications for the thousands of kinds of protein molecules the cell requires for its moment-to-moment existence: for extracting energy from molecules assimilated as food and for repairing itself as well as for replication. It is only within the past 15 years, however, that insight has been gained into the chemical nature of the genetic material and how its molecular structure can embody coded instructions that can be "read" by the machinery in the cell responsible for synthesizing protein molecules. As the result of intensive work by many investigators the story



SYNTHESIS OF PROTEIN MOLECULES is accomplished by the intracellular particles called ribosomes. The coded instructions for making the protein molecule are carried to the ribosome by a form of ribonucleic acid (RNA) known as "messenger" RNA. The RNA code "letters" are four bases: uracil (U), cytosine (C), adenine (A) and guanine (G). A sequence of three bases, called a codon, is required to specify each of the 20 kinds of amino acid, identified here by their abbreviations. (A list of the 20 amino acids and their abbreviations appears on the next page.) When linked end to end, these amino acids form the polypeptide chains of which proteins are composed. Each type of amino acid is transported to the ribosome by a particular form of "transfer" RNA (tRNA), which carries an anticodon that can form a temporary bond with one of the codons in messenger RNA. Here the ribosome is shown moving along the chain of messenger RNA, "reading off" the codons in sequence. It appears that the ribosome has two binding sites for molecules of tRNA: one site (*A*) for positioning a newly arrived tRNA molecule and another (*B*) for holding the growing polypeptide chain.

AMINO ACID	ABBREVIATION
ALANINE	Ala
ARGININE	Arg
ASPARAGINE	AspN
ASPARTIC ACID	Asp
CYSTEINE	Cys
GLUTAMIC ACID	Glu
GLUTAMINE	GluN
GLYCINE	Gly
HISTIDINE	His
ISOLEUCINE	Ileu
LEUCINE	Leu
LYSINE	Lys
METHIONINE	Met
PHENYLALANINE	Phe
PROLINE	Pro
SERINE	Ser
THREONINE	Thr
TRYPTOPHAN	Tryp
TYROSINE	Tyr
VALINE	Val

TWENTY AMINO ACIDS constitute the standard set found in all proteins. A few other amino acids occur infrequently in proteins but it is suspected in each case that they originate as one of the standard set and become chemically modified after they have been incorporated into a polypeptide chain.

of the genetic code is now essentially complete. One can trace the transmission of the coded message from its original site in the genetic material to the finished protein molecule.

The genetic material of the living cell is the chainlike molecule of deoxyribonucleic acid (DNA). The cells of many bacteria have only a single chain; the cells of mammals have dozens clustered together in chromosomes. The DNA molecules have a very long backbone made up of repeating groups of phosphate and a five-carbon sugar. To this backbone the side groups called bases are attached at regular intervals. There are four standard bases: adenine (A), guanine (G), thymine (T) and cytosine (C). They are the four "letters" used to spell out the genetic message. The exact sequence of bases along a length of the DNA molecule determines the structure of a particular protein molecule.

Proteins are synthesized from a standard set of 20 amino acids, uniform throughout nature, that are joined end to end to form the long polypeptide chains of protein molecules [*see illustration at left*]. Each protein has its own characteristic sequence of amino acids. The number of amino acids in a polypeptide chain ranges typically from 100 to 300 or more.

The genetic code is not the message itself but the "dictionary" used by the cell to translate from the four-letter language of nucleic acid to the 20-letter language of protein. The machinery of the cell can translate in one direction only: from nucleic acid to protein but not from protein to nucleic acid. In making this translation the cell employs a variety of accessory molecules and mechanisms. The message contained in DNA is first transcribed into the similar molecule called "messenger" ribonucleic acid—messenger RNA. (In many viruses—the tobacco mosaic virus, for example—the genetic material is simply RNA.) RNA too has four kinds of bases as side groups; three are identical with those found in DNA (adenine, guanine and cytosine) but the fourth is uracil (U) instead of thymine. In this first transcription of the genetic message the code letters A, G, T and C in DNA give rise respectively to U, C, A and G. In other words, wherever A appears in DNA, U appears in the RNA transcription; wherever G appears in DNA, C appears in the transcription, and so on. As it is usually presented the dictionary of the genetic code employs the letters found in RNA (U, C, A, G) rather than those found in DNA (A, G, T, C).

The genetic code could be broken easily if one could determine both the amino acid sequence of a protein and the base sequence of the piece of nucleic acid that codes it. A simple comparison of the two sequences would yield the code. Unfortunately the determination of the base sequence of a long nucleic acid molecule is, for a variety of reasons, still extremely difficult. More indirect approaches must be used.

Most of the genetic code first became known early in 1965. Since then additional evidence has proved that almost all of it is correct, although a few features remain uncertain. This article describes how the code was discovered and some of the work that supports it.

Scientific American has already presented a number of articles on the genetic code. In one of them ["The Genetic Code," October, 1962] I explained that the experimental evidence (mainly indirect) suggested that the code was a triplet code: that the bases on the messenger RNA were read three at a time and that each group corresponded to a particular amino acid. Such a group is called a codon. Using four symbols in groups of three, one can form 64 distinct triplets. The evidence indicated that most of these stood for one amino acid or another, implying that an amino acid was usually represented by several codons. Adjacent amino acids were coded by adjacent codons, which did not overlap.

In a sequel to that article ["The Genetic Code: II," March, 1963] Marshall W. Nirenberg of the National Institutes of Health explained how the composition of many of the 64 triplets had been determined by actual experiment. The technique was to synthesize polypeptide chains in a cell-free system, which was made by breaking open cells of the colon bacillus (*Escherichia coli*) and extracting from them the machinery for protein synthesis. Then the system was provided with an energy supply, 20 amino acids and one or another of several types of synthetic RNA. Although the exact sequence of bases in each type was random, the proportion of bases was known. It was found that each type of synthetic messenger RNA directed the incorporation of certain amino acids only.

By means of this method, used in a quantitative way, the *composition* of many of the codons was obtained, but the *order* of bases in any triplet could not be determined. Codons rich in G were difficult to study, and in addition a few mistakes crept in. Of the 40 codon compositions listed by Nirenberg in his article we now know that 35 were correct.

The Triplet Code

The main outlines of the genetic code were elucidated by another technique invented by Nirenberg and Philip Leder. In this method no protein synthesis occurs. Instead one triplet at a time is used to bind together parts of the machinery of protein synthesis.

Protein synthesis takes place on the comparatively large intracellular structures known as ribosomes. These bodies travel along the chain of messenger RNA, reading off its triplets one after another and synthesizing the polypeptide chain of the protein, starting at the amino end (NH_2). The amino acids do not diffuse to the ribosomes by themselves. Each amino acid is joined chemically by a special enzyme to one of the codon-recognizing molecules known both as soluble RNA (sRNA) and transfer RNA (tRNA). (I prefer the latter designation.) Each tRNA mole-

cule has its own triplet of bases, called an anticodon, that recognizes the relevant codon on the messenger RNA by pairing bases with it [*see illustration on page 55*].

Leder and Nirenberg studied which amino acid, joined to its tRNA molecules, was bound to the ribosomes in the presence of a particular triplet, that is, by a "message" with just three letters. They did so by the neat trick of passing the mixture over a nitrocellulose filter that retained the ribosomes. All the tRNA molecules passed through the filter except the ones specifically bound to the ribosomes by the triplet. Which they were could easily be decided by using mixtures of amino acids in which one kind of amino acid had been made artificially radioactive, and determining the amount of radioactivity absorbed by the filter.

For example, the triplet GUU retained the tRNA for the amino acid valine, whereas the triplets UGU and UUG did not. (Here GUU actually stands for the trinucleoside diphosphate GpUpU.) Further experiments showed that UGU coded for cysteine and UUG for leucine.

Nirenberg and his colleagues synthesized all 64 triplets and tested them for their coding properties. Similar results have been obtained by H. Gobind Khorana and his co-workers at the University of Wisconsin. Various other groups have checked a smaller number of codon assignments.

Close to 50 of the 64 triplets give a clearly unambiguous answer in the binding test. Of the remainder some evince only weak binding and some bind more than one kind of amino acid. Other results I shall describe later suggest that the multiple binding is often an artifact of the binding method. In short, the binding test gives the meaning of the majority of the triplets but it does not firmly establish all of them.

The genetic code obtained in this way, with a few additions secured by other methods, is shown in the table below. The 64 possible triplets are set out in a regular array, following a plan

SECOND LETTER

FIRST LETTER	U	C	A	G	THIRD LETTER
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA OCHRE	UGA ?	A
	UUG Leu	UCG Ser	UAG AMBER	UGG Tryp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA GluN	CGA Arg	A
	CUG Leu	CCG Pro	CAG GluN	CGG Arg	G
A	AUU Ileu	ACU Thr	AAU AspN	AGU Ser	U
	AUC Ileu	ACC Thr	AAC AspN	AGC Ser	C
	AUA Ileu	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

GENETIC CODE, consisting of 64 triplet combinations and their corresponding amino acids, is shown in its most likely version. The importance of the first two letters in each triplet is readily apparent. Some of the allocations are still not completely certain, particularly for organisms other than the colon bacillus (*Escherichia coli*). "Amber" and "ochre" are terms that referred originally to certain mutant strains of bacteria. They designate two triplets, UAA and UAG, that may act as signals for terminating polypeptide chains.

that clarifies the relations between them.

Inspection of the table will show that the triplets coding for the same amino acid are often rather similar. For example, all four of the triplets starting with the doublet AC code for threonine. This pattern also holds for seven of the other amino acids. In every case the triplets *XYU* and *XYC* code for the same amino acid, and in many cases *XYA* and *XYG* are the same (methionine and tryptophan may be exceptions). Thus an amino acid is largely selected by the first two bases of the triplet. Given that a triplet codes for, say, valine, we know that the first two bases are GU, whatever the third may be. This pattern is true for all but three of the amino acids. Leucine can start with UU or CU, serine with UC or AG and arginine with CG or AG. In all other cases the amino acid is uniquely related to the first two bases of the triplet. Of course, the converse is often not true. Given that a triplet starts with, say, CA, it may code for either histidine or glutamine.

Synthetic Messenger RNA's

Probably the most direct way to confirm the genetic code is to synthesize a messenger RNA molecule with a strictly defined base sequence and then find the amino acid sequence of the polypeptide produced under its influence. The most extensive work of this nature has been done by Khorana and his colleagues. By a brilliant combination of ordinary chemical synthesis and synthesis catalyzed by enzymes, they have made long RNA molecules with various repeating sequences of bases. As an example, one RNA molecule they have synthesized has the sequence UGUG-UGUGUGUG.... When the biochemical machinery reads this as triplets the message is UGU–GUG–UGU–GUG.... Thus we expect that a polypeptide will be produced with an alternating sequence of two amino acids. In fact, it was found that the product is Cys–Val–Cys–Val.... This evidence alone would not tell us which triplet goes with which amino acid, but given the results of the binding test one has no hesitation in concluding that UGU codes for cysteine and GUC for valine.

In the same way Khorana has made chains with repeating sequences of the type *XYZ*... and also *XXYZ*.... The type *XYZ*... would be expected to give a "homopolypeptide" containing one amino acid corresponding to the triplet *XYZ*. Because the starting point is not clearly defined, however, the homopolypeptides corresponding to *YZX*... and *ZXY*... will also be produced. Thus poly-AUC makes polyisoleucine, polyserine and polyhistidine. This confirms that AUC codes for isoleucine, UCA for serine and CAU for histidine. A repeating sequence of four bases will yield a single type of polypeptide with a repeating sequence of four amino acids. The general patterns to be expected in each case are set forth in the table on this page. The results to date have amply demonstrated by a direct biochemical method that the code is indeed a triplet code.

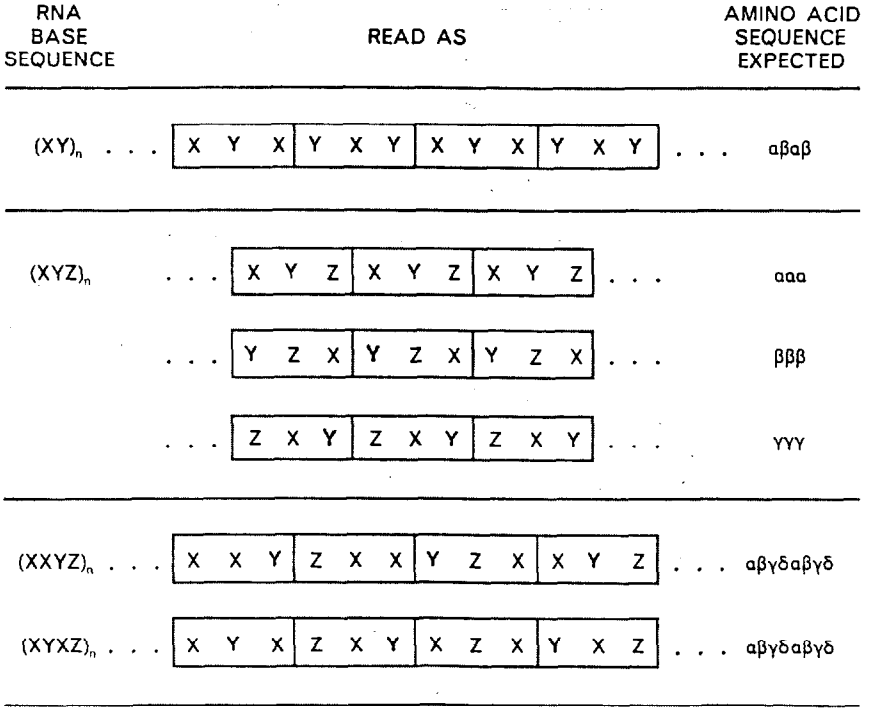
Khorana and his colleagues have so far confirmed about 25 triplets by this method, including several that were quite doubtful on the basis of the binding test. They plan to synthesize other sequences, so that eventually most of the triplets will be checked in this way.

The Use of Mutations

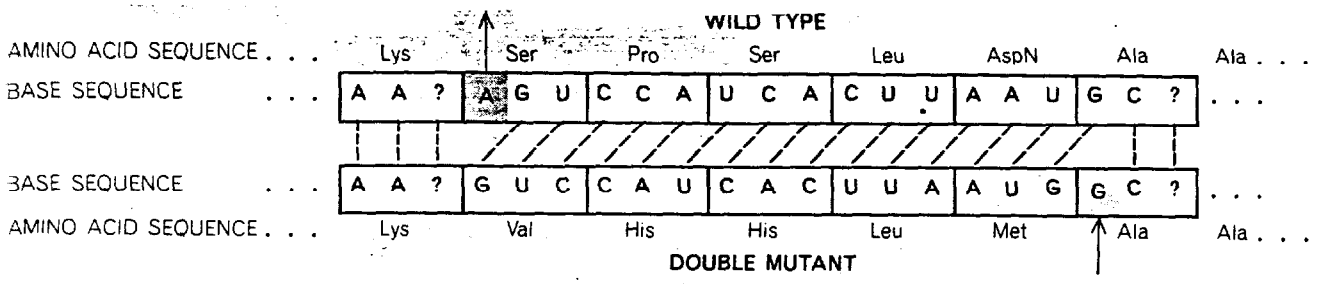
The two methods described so far are open to the objection that since they do not involve intact cells there may be some danger of false results. This objection can be met by two other methods of checking the code in which the act of protein synthesis takes place inside the cell. Both involve the effects of genetic mutations on the amino acid sequence of a protein.

It is now known that small mutations are normally of two types: "base substitution" mutants and "phase shift" mutants. In the first type one base is changed into another base but the total number of bases remains the same. In the second, one or a small number of bases are added to the message or subtracted from it.

There are now extensive data on base-substitution mutants, mainly from studies of three rather convenient proteins: human hemoglobin, the protein of tobacco mosaic virus and the *A* protein of the enzyme tryptophan synthetase obtained from the colon bacillus. At least 36 abnormal types of human hemoglobin have now been investigated by many different workers. More than 40 mutant forms of the protein of the tobacco mosaic virus have been examined by Hans Wittmann of the Max Planck Institute for Molecular Genetics in Tübingen and by Akita Tsugita and Heinz Fraenkel-Conrat of the University of California at Berkeley [see "The Genetic Code of a Virus," by Heinz Fraenkel-Conrat; SCIENTIFIC AMERICAN, October, 1964]. Charles Yanofsky and his group at Stanford University have characterized about 25 different mutations of the *A* protein of tryptophan synthetase.



VARIETY OF SYNTHETIC RNA's with repeating sequences of bases have been produced by H. Gobind Khorana and his colleagues at the University of Wisconsin. They contain two or three different bases (X, Y, Z) in groups of two, three or four. When introduced into cell-free systems containing the machinery for protein synthesis, the base sequences are read off as triplets (*middle*) and yield the amino acid sequences indicated at the right.



"PHASE SHIFT" MUTATIONS help to establish the actual codons used by organisms in the synthesis of protein. The two partial amino acid sequences shown here were determined by George Streisinger and his colleagues at the University of Oregon. The sequences are from a protein, a type of lysozyme, produced by the bacterial virus T4. A pair of phase-shift mutations evidently removed one base, A, and inserted another, G, about 15 bases farther on. The base sequence was deduced theoretically from the genetic code.

The remarkable fact has emerged that in every case but one the genetic code shows that the change of an amino acid in a polypeptide chain could have been caused by the alteration of a single base in the relevant nucleic acid. For example, the first observed change of an amino acid by mutation (in the hemoglobin of a person suffering from sickle-cell anemia) was from glutamic acid to valine. From the genetic code dictionary on page 57 we see that this could have resulted from a mutation that changed either GAA to GUA or GAG to GUG. In either case the change involved a single base in the several hundred needed to code for one of the two kinds of chain in hemoglobin.

The one exception so far to the rule that all amino acid changes could be caused by single base changes has been found by Yanofsky. In this one case glutamic acid was replaced by methionine. It can be seen from the genetic code dictionary that this can be accomplished only by a change of *two* bases, since glutamic acid is encoded by either GAA or GAG and methionine is encoded only by AUG. This mutation has occurred only once, however, and of all the mutations studied by Yanofsky it is the only one not to back-mutate, or revert to "wild type." It is thus almost certainly the rare case of a double change. All the other cases fit the hypothesis that base-substitution mutations are normally caused by a single base change. Examination of the code shows that only about 40 percent of all the possible amino acid interchanges can be brought about by single base substitutions, and it is only these changes that are found in experiments. Therefore the study of actual mutations has provided strong confirmation of many features of the genetic code.

Because in general several codons stand for one amino acid it is not possible, knowing the amino acid sequence, to write down the exact RNA base sequence that encoded it. This is unfortunate. If we know which amino acid is changed into another by mutation, however, we can often, given the code, work out what that base change must have been. As an example, glutamic acid can be encoded by GAA or GAG and valine by GUU, GUC, GUA or GUG. If a mutation substitutes valine for glutamic acid, one can assume that only a single base change was involved. The only such change that could lead to the desired result would be a change from A to U in the middle position, and this would be true whether GAA became GUA or GAG became GUG.

It is thus possible in many cases (not in all) to compare the nature of the base change with the chemical mutagen used to produce the change. If RNA is treated with nitrous acid, C is changed to U and A is effectively changed to G. On the other hand, if double-strand DNA is treated under the right conditions with hydroxylamine, the mutagen acts only on C. As a result some C's are changed to T's (the DNA equivalent of U's), and thus G's, which are normally paired with C's in double-strand DNA, are replaced by A's.

If 2-aminopurine, a "base analogue" mutagen, is added when double-strand DNA is undergoing replication, it produces only "transitions." These are the same changes as those produced by hydroxylamine—plus the reverse changes. In almost all these different cases (the exceptions are unimportant) the changes observed are those expected from our knowledge of the genetic code.

Note the remarkable fact that, although the code was deduced mainly from studies of the colon bacillus, it appears to apply equally to human beings and tobacco plants. This, together with more fragmentary evidence, suggests that the genetic code is either the same or very similar in most organisms.

The second method of checking the code using intact cells depends on phase-shift mutations such as the addition of a single base to the message. Phase-shift mutations probably result from errors produced during genetic recombination or when the DNA molecule is being duplicated. Such errors have the effect of putting out of phase the reading of the message from that point on. This hypothesis leads to the prediction that the phase can be corrected if at some subsequent point a nucleotide is deleted. The pair of alterations would be expected not only to change two amino acids but also to alter all those encoded by bases lying between the two affected sites. The reason is that the intervening bases would be read out of phase and therefore grouped into triplets different from those contained in the normal message.

This expectation has recently been confirmed by George Streisinger and his colleagues at the University of Oregon. They have studied mutations in the protein lysozyme that were produced by the T4 virus, which infects the colon bacillus. One phase-shift mutation involved the amino acid sequence ...Lys–Ser–Pro–Ser–Leu–AspN–Ala–Ala–Lys.... They were then able to construct by genetic methods a double phase-shift mutant in which the corresponding sequence was ...Lys–Val–His–His–Leu–Met–Ala–Ala–Lys....

Given these two sequences, the reader should be able, using the genetic code dictionary on page 57, to decipher uniquely a short length of the nucleic acid message for both the original protein and the double mutant and thus deduce the changes produced by each of the phase-shift mutations. The correct result is presented in the illustration above. The result not only confirms several rather doubtful codons, such as UUA for leucine and AGU for serine, but also shows which codons are actually involved in a genetic message. Since the technique is difficult, however, it may not find wide application.

Streisinger's work also demonstrates what has so far been only tacitly as-

ANTICODON	CODON
U	A G
C	G
A	U
G	U C
I	U C A

"WOBBLE" HYPOTHESIS has been proposed by the author to provide rules for the pairing of codon and anticodon at the *third* position of the codon. There is evidence, for example, that the anticodon base I, which stands for inosine, may pair with as many as three different bases: U, C and A. Inosine closely resembles the base guanine (G) and so would ordinarily be expected to pair with cytosine (C). Structural diagrams for standard base pairings and wobble base pairings are illustrated at the bottom of this page.

sumed: that the two languages, both of which are written down in a certain direction according to convention, are in fact translated by the cell in the same direction and not in opposite directions. This fact had previously been established, with more direct chemical methods, by Severo Ochoa and his colleagues at the New York University School of Medicine. In the convention, which was adopted by chance, proteins are written with the amino (NH_2) end on the left. Nucleic acids are written with the end of the molecule containing a 5-prime carbon atom at the left. (The "5 prime" refers to a particular carbon atom in the 5-carbon ring of ribose sugar or deoxyribose sugar.)

Finding the Anticodons

Still another method of checking the genetic code is to discover the three bases making up the anticodon in some particular variety of transfer RNA. The first tRNA to have its entire sequence worked out was alanine tRNA, a job done by Robert W. Holley and his collaborators at Cornell University [see "The Nucleotide Sequence of a Nucleic Acid," by Robert W. Holley; SCIENTIFIC AMERICAN, February]. Alanine tRNA, obtained from yeast, contains 77 bases. A possible anticodon found near the middle of the molecule has the sequence IGC, where I stands for inosine, a base closely resembling guanine. Since then Hans Zachau and his colleagues at the University of Cologne have established the sequences of two closely related serine tRNA's from yeast, and James Madison and his group at the U.S. Plant, Soil and Nutrition Laboratory at Ithaca, N.Y., have worked out the sequence of a tyrosine tRNA, also from yeast.

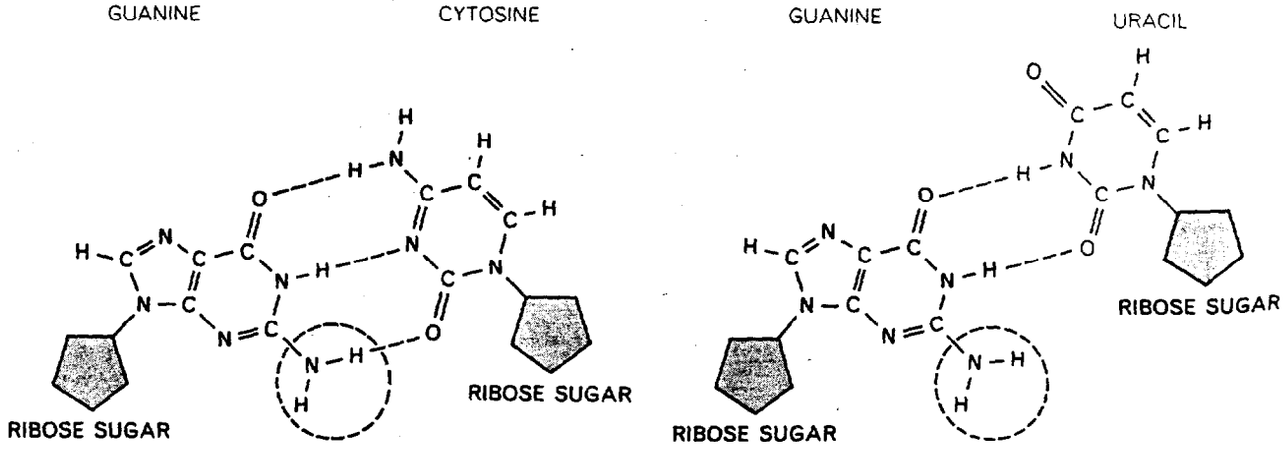
A detailed comparison of these three sequences makes it almost certain that the anticodons are alanine–IGC, serine–IGA and tyrosine–GΨA. (Ψ stands for pseudo-uridylic acid, which can form the same base pairs as the base uracil.) In addition there is preliminary evidence from other workers that an anticodon for valine is IAC and an anticodon for phenylalanine is GAA.

All these results would fit the rule that the codon and anticodon pair in an antiparallel manner, and that the pairing in the first two positions of the codon is of the standard type, that is, A pairs with U and G pairs with C. The pairing in the third position of the codon is more complicated. There is now good experimental evidence from both Nirenberg and Khorana and their co-workers that one tRNA can recognize several codons, provided that they differ only in the last place in the codon. Thus Holley's alanine tRNA appears to recognize GCU, GCC and GCA. If it recognizes GCG, it does so only very weakly.

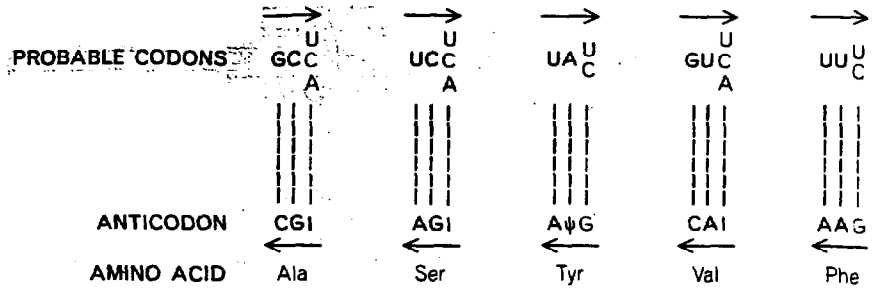
The "Wobble" Hypothesis

I have suggested that this is because of a "wobble" in the pairing in the third place and have shown that a reasonable theoretical model will explain many of the observed results. The suggested rules for the pairing in the third position of the anticodon are presented in the table at the top of this page, but this theory is still speculative. The rules for the first two places of the codon seem reasonably secure, however, and can be used as partial confirmation of the genetic code. The likely codon-anticodon pairings for valine, serine, tyrosine, alanine and phenylalanine satisfy the standard base pairings in the first two places and the wobble hypothesis in the third place [*see illustration on page 62*].

Several points about the genetic code remain to be cleared up. For example, the triplet UGA has still to be allocated.



STANDARD AND WOBBLE BASE PAIRINGS both involve the formation of hydrogen bonds when certain bases are brought into close proximity. In the standard guanine-cytosine pairing (*left*) it is believed three hydrogen bonds are formed. The bases are shown as they exist in the RNA molecule, where they are attached to 5-carbon rings of ribose sugar. In the proposed wobble pairing (*right*) guanine is linked to uracil by only two hydrogen bonds. The base inosine (I) has a single hydrogen atom where guanine has an amino (NH_2) group (*broken circle*). In the author's wobble hypothesis inosine can pair with U as well as with C and A (*not shown*).



CODON-ANTICODON PAIRINGS take place in an antiparallel direction. Thus the anticodons are shown here written backward, as opposed to the way they appear in the text. The five anticodons are those tentatively identified in the transfer RNA's for alanine, serine, tyrosine, valine and phenylalanine. Color indicates where wobble-pairings may occur.

The punctuation marks—the signals for "begin chain" and "end chain"—are only partly understood. It seems likely that both the triplet UAA (called "ochre") and UAG (called "amber") can terminate the polypeptide chain, but which triplet is normally found at the end of a gene is still uncertain.

The picturesque terms for these two triplets originated when it was discovered in studies of the colon bacillus some years ago that mutations in other genes (mutations that in fact cause errors in chain termination) could "suppress" the action of certain mutant codons, now identified as either UAA or UAG. The terms "ochre" and "amber" are simply invented designations and have no reference to color.

A mechanism for chain initiation was discovered fairly recently. In the colon bacillus it seems certain that formylmethionine, carried by a special tRNA, can initiate chains, although it is not clear if all chains have to start in this way, or what the mechanism is in mammals and other species. The formyl group (CHO) is not normally found on finished proteins, suggesting that it is probably removed by a special enzyme. It seems likely that sometimes the methionine is removed as well.

It is unfortunately possible that a few codons may be ambiguous, that is, may code for more than one amino acid. This is certainly not true of most codons. The present evidence for a small amount of ambiguity is suggestive but not conclusive. It will make the code more difficult to establish correctly if ambiguity can occur.

Problems for the Future

From what has been said it is clear that, although the entire genetic code is not known with complete certainty, it is highly likely that most of it is correct. Further work will surely clear up the doubtful codons, clarify the punctuation marks, delimit ambiguity and extend the code to many other species. Although the code lists the codons that *may* be used, we still have to determine if alternative codons are used equally. Some preliminary work suggests they may not be. There is also still much to be discovered about the machinery of protein synthesis. How many types of tRNA are there? What is the structure of the ribosome? How does it work, and why is it in two parts? In addition there are many questions concerning the control of the rate of protein synthesis that we are still a long way from answering.

When such questions have been answered, the major unsolved problem will be the structure of the genetic code. Is the present code merely the result of a series of evolutionary accidents, so that the allocations of triplets to amino acids is to some extent arbitrary? Or are there profound structural reasons why phenylalanine has to be coded by UUU and UUC and by no other triplets? Such questions will be difficult to decide, since the genetic code originated at least three billion years ago, and it may be impossible to reconstruct the sequence of events that took place at such a remote period. The origin of the code is very close to the origin of life. Unless we are lucky it is likely that much of the evidence we should like to have has long since disappeared.

Nevertheless, the genetic code is a major milestone on the long road of molecular biology. In showing in detail how the four-letter language of nucleic acid controls the 20-letter language of protein it confirms the central theme of molecular biology that genetic information can be stored as a one-dimensional message on nucleic acid and be expressed as the one-dimensional amino acid sequence of a protein. Many problems remain, but this knowledge is now secure.